Composition and Regulation of Maternal and Zygotic Transcriptomes Reflects Species Specific Reproductive Mode

Composition and regulation of maternal and zygotic transcriptomes reflects species specific reproductive mode

Shai S Shen-Orr[1†], Yitzhak Pilpel[2], Craig P Hunter[1§]

[1] Department of Molecular and Cellular Biology, Harvard University, 16 Divinity Ave., Cambridge, MA, 02138, USA

[2]Department of Molecular Genetics, Weizmann Institute of Science, Rehovot, 76100, Israel

[†] Current address: Departments of Pediatrics and Microbiology & Immunology, Stanford University, Stanford, CA 94305, USA

[§]Corresponding author

Email addresses:

SSO: shenorr@stanford.edu

YP: pilpel@weizmann.ac.il

CPH: hunter@mcb.harvard.edu

Abstract

Background

Early embryos contain mRNA transcripts expressed from two distinct origins; those expressed from the mother's genome and deposited in the oocyte (maternal) and those expressed from the embryo's genome after fertilization (zygotic). The transition from maternal to zygotic control occurs at different times in different animals according to the extent and form of maternal contributions, which likely reflect evolutionary and ecological forces. Maternally deposited transcripts rely on post-transcriptional regulatory mechanisms for precise spatial and temporal expression in the embryo, whereas zygotic transcripts can use both transcriptional and post-transcriptional regulatory mechanisms. The differences in maternal contributions between animals may be associated with gene regulatory changes detectable by the size and complexity of the associated regulatory regions.

Results

We have used genomic data to identify and compare maternal and/or zygotic expressed genes from six different animals and find evidence for selection acting to shape gene regulatory architecture in thousands of genes. We find that mammalian maternal genes are enriched for complex regulatory regions suggesting an increase in expression specificity, while egg-laying animals are enriched for maternal genes that lack transcriptional specificity.

Conclusions

We propose that this lack of specificity for maternal expression in egg-laying animals indicates that a large fraction of maternal genes are expressed non-functionally, providing only supplemental nutritional content to the developing embryo. These results provide clear predictive criteria for analysis of additional genomes.

Background

Early embryos contain mRNA transcripts expressed from two distinct origins; those expressed from the mother’s genome and deposited in the oocyte (maternal) and those expressed from the embryo’s genome after fertilization (zygotic). Because these

transcripts originate from distinct origins they are subject to distinct regulatory constraints. Maternal transcripts rely on post-transcriptional regulatory mechanisms for spatial and temporal control of their embryonic expression, and thus contain all signals that control their stability, localization and relative accessibility to the translational machinery [1-7]. In contrast zygotically synthesized transcripts may utilize both transcriptional and post-transcriptional regulatory mechanisms to provide precise temporal and spatial expression.

In all animals surveyed to date, at least 30% of protein-coding genes are detected as expressed during the transition from unfertilized oocyte to early embryo [8-13]. These may be divided into three basic groups. First, those that must be expressed exclusively from either a maternal or a zygotic origin, which include maternally expressed genes required to "jump-start" embryogenesis and zygotically expressed patterning genes whose precocious (maternal) expression would disrupt temporal or spatial developmental events [14]. Second, those which must be expressed both by the mother and by the embryo, for example because of low mRNA stability or because of a change in spatial expression in transition between oocyte and embryo [15]. The last group is those genes that can accommodate either maternal or zygotic expression. It is among this latter gene set that evolution can act to maximize the efficiency, or other such measure, of embryogenesis or oogenesis.

A gene's regulatory architecture reflects the extent and complexity of transcriptional and post-transcriptional gene expression. For example, a gene such as the sea urchin *endo-16*, which is subject to complex spatial and temporal regulation at a multi-cellular stage of embryogenesis contains a large complex intergenic regulatory region [16]. In contrast, a gene such as Drosophila Oskar, which is transcribed maternally and subject to multiple levels of post-transcriptional regulation has a large 3' untranslated region (UTR) that controls transcript localization, stability, and translation [17]. Finally, many house-keeping genes are ubiquitously expressed and consequently have relatively simple regulatory needs.

At present, accurately and comprehensively assessing the regulatory architecture of the majority of genes is difficult, as the regulation of only a few has been well-characterized [18]. Yet, in organisms with relatively small genomes (up to 150MB), genes expressed in many tissues or involved in complex biological processes have longer than average 5' intergenic regions (IGRs) [19, 20] and 3' UTRs [21]. Furthermore, the sizes of these regulatory regions correlate positively with the number of known and/or predicted *cis*-regulatory sites [20-22]. Particularly interesting in the context of our study is the observation that the 3' UTRs of maternal genes in *D. melanogaster* are longer than average, suggesting that they are subject to greater post-transcriptional control [5].

In organisms with larger genomes, such as human, housekeeping genes are flanked by small intergenic regions (IGRs) [23-25] and are associated with low density of conserved non-coding elements. Conversely, genes neighboring large gene-free regions or having large introns have dense regulatory elements and are associated with developmental functions and tissue specificity [25-27]. To first principles, these observations provide a means to assess a gene regulatory architecture, where the extent of regulation is approximated by the length of the regulatory regions, and the type of the region, IGR or UTR, identifies whether the regulation is respectively transcriptional or post-transcriptional.

Here, we assess the differing regulatory constraints between maternal and zygotically expressed genes by analyzing the regulatory architecture of individual genes. To do so, we used mRNA time-course expression data to identify maternal and zygotic genes in worm, fly, fish and mouse (*C. elegans, D. melanogaster, D. rerio and M. musculus*). For each data set, at least one time point was collected prior to the start of major zygotic transcription, and at least one time point after [4, 9, 10, 15]. In addition, genome-wide mRNA expression datasets from chicken (*G. gallus*) eggs and human oocytes allowed identification of maternally expressed genes in those organisms [12, 28]. Comparative analysis of maternal and zygotic genes within an animal reveals the effect of a yet un-described selective evolutionary forces acting to modify the gene regulatory architecture of thousands of genes, as a function of germline versus embryonic transcript synthesis. In

contrast, cross-species comparisons allow studying this force and understanding the factors that affect it. These show that this selective force affecting gene regulation at the molecular level is in agreement to the alternative strategies for managing maternal versus zygotic energy expenditures at the physiological level, suggesting the maintenance of a delicate balance between different energy resources utilized to “jump-start” embryonic development.

Results

Across the animal kingdom, 3’ UTRs of maternally expressed genes are not short, reflecting the requirement for post-transcriptional regulation of maternal genes

Genes whose transcripts were detected as present in the embryo before the initiation of zygotic transcription were defined as members of the ”all-maternal” gene class (see Materials and Methods). To compare the relative contribution of post-transcriptional regulation among different classes of maternal transcripts we used the length of the 3’ UTR as an estimate of the complexity of a gene’s post-transcriptional program (addition of 5’UTR length yielded qualitatively similar results, see Materials and Methods). To account for differences in functional complexity [19-21, 26, 29], we applied a genome-wide phylogenetic profile of 26 organisms [30] to classify genes as either "core", conserved in both uni-cellular and multi-cellular organisms, or as "metazoan", and analyzed them separately. In all animals the 3’ UTR lengths of the all-maternal class genes were significantly under-represented for short lengths compared to all other coding genes (Figure 1a,b, p-value < 0.05 in all cases using a modified Kolmogorov-Smirnov test; see figure legend and Materials and Methods for details). In addition, with the exception of *C. elegans* and *G. gallus,* significant differences were also detected between all-maternal core and metazoan genes. This preservation of 3’ UTR length among maternal transcripts occurs across a 30-fold range in genome size (100MB - 3GB), a 5-fold range in genome-wide mean 3’ UTR length (150bp - 900bp), and large differences in development and stability of maternal transcripts [7, 31, 32]. We conclude that across the animal kingdom the post-transcriptional regulatory constraint imposed on maternally expressed genes has selected against short 3’ UTRs.

D. melanogaster zygotic genes have longer 5' IGRs whereas maternal genes are under-represented for short 3' UTRs

After the initiation of zygotic transcription, the assignment of relative maternal and zygotic transcription to a gene's measured mRNA abundance becomes less certain. However, for *D. melanogaster*, exact quantification of relative maternal and zygotic contributions to mRNA abundance was made possible through the use of embryos lacking entire chromosomes [15]. This analysis defined five separate gene classes for transcripts detected in early embryos (see Materials and Methods): strict-maternal and strict-zygotic genes are expressed solely from one origin of expression; mostly-maternal and mostly-zygotic genes are those whose expression profile is similar to their strict counterparts, but for whom at least some contribution (less than 33%) is due to zygotic or maternal origin respectively [15]; and finally, the maternal-zygotic genes are those which are transcribed maternally, but whose transcript abundance level does not change significantly throughout the duration of the experiment (either stable or supplemented by zygotic transcription).

Comparison of 3' UTR lengths between the five different origin-of-synthesis classes showcases the effect of the biological constraints on 3' UTR length. The 3' UTRs of maternal and zygotic class genes are significantly longer than those of other genes in the genome. In particular, with the exception of the core strict-zygotic class, both core and metazoan strict-maternal genes are underrepresented for short 3' UTRs compared to all other classes (Figure S1, across all comparisons p-value at least <= 0.02). Interestingly the longest 3' UTRs are those of zygotic genes.

Significant differences are also observed between maternal and zygotic genes with respect to 5' IGR lengths (addition of intron lengths and/or 3'IGR lengths yielded qualitatively similar results, see Material and Methods). For metazoan genes, the four gene classes that include some maternally contributed transcripts, have significantly shorter 5' IGR lengths than all other metazoan genes in the genome (Figure 2a) ($p<10^{-9}$, $p<10^{-4}$, $p<10^{-12}$, $p<10^{-5}$ for strict-maternal, mostly-maternal, maternal-zygotic and mostly-zygotic respectively). Strikingly, the 5' IGR lengths of the small set of 282 genes belonging to the strictly-zygotic class are extremely long compared to all other gene sets (Strict-maternal [Core: $p<10^{-5}$, Metazoan: $p<10^{-18}$]; mostly-maternal [$p<10^{-6}$, $p<10^{-12}$];

mostly-maternal [$p<10^{-7}$, $p<10^{-18}$]; mostly-zygotic [$p<10^{-6}$, $p<10^{-13}$]; and the genome-wide set of all core and metazoan genes [$p<10^{-11}$, $p<10^{-10}$]). Interestingly, this class is enriched for patterning genes ($p<10^{-32}$ De Renzis et al. [15]), whereas the strict-maternal class is enriched for core genes ($p<10^{-115}$), as would be expected from the proposed theory on maternal and zygotic gene expression in rapidly developing organisms [14]. Lastly, comparing the core genes to metazoan genes the 3' UTRs and 5' IGRs of core genes are shorter for nearly all maternal and zygotic classes (Strict-maternal [3' UTR: $p<10^{-6}$, 5' IGR: $p<0.07$]; mostly-maternal [$p<10^{-9}$, $p<10^{-6}$]; maternal-zygotic [$p<10^{-35}$, $p<10^{-21}$]; mostly-zygotic [$p<10^{-12}$, $p<10^{-7}$]; strictly-zygotic [$p<10^{-4}$, $p<10^{-3}$]; and the genome-wide set of all core and metazoan genes [$p<10^{-21}$, $p<10^{-72}$]).

Similarity in regulatory architecture of maternal and zygotic genes across the animal kingdom, highlights the complexity of regulation of mammalian maternal genes

To analyze the gene architecture of maternal and zygotically expressed genes in other animals (*C. elegans*, *D. rerio G. gallus*, *M. musculus* and *H. sapiens*) we defined three gene classes for transcripts detected in early embryos; maternal, zygotic and maternal-zygotic. For chicken and human, to the best of our knowledge, only pre-zygotic transcript data is publicly available, thus for these species we contrasted the all-maternal gene class with the genome-wide set of core and metazoan genes. Further, due to the lack of genetic controls available in Drosophila, for these other species we must rely on the characteristic expression profile to define the origin of expression (see Materials and Methods). For clarity, we use the nomenclature applied to the Drosophila data and refer to the maternal and zygotic gene classes as strict-maternal and strict-zygotic. By necessity, the maternal-zygotic class is less precisely defined and includes slowly decaying strict-maternal genes. Consistent with this, we find that the lengths of the regulatory regions in the maternal-zygotic class are by and large intermediate to those observed in the strict-maternal and strict-zygotic gene classes (data not shown). Therefore, unless otherwise noted, we exclude the maternal-zygotic class from further analysis.

Next, for each species we compared the 5' IGRs lengths as proxies for the functional complexity of maternal and zygotic gene regulatory regions. Additionally, within these origin-of-synthesis class gene sets, we compared the core and metazoan subclasses to the genome-wide core or metazoan gene sets (see Materials and Methods). Because it is meaningless to compare the absolute lengths of genes' regulatory region size across species with vastly different genome sizes, the genome-wide core or metazoan gene sets provide a means to normalize length for cross species comparisons. Performing this comparative analysis between maternal and zygotic gene classes separates the studied animals into two distinct groups. *C. elegans, D. rerio* and *G. gallus* genes show a pattern similar to that described for *D. melanogaster*. The 5' IGRs of *C. elegans* and *D. rerio* strict-maternal genes (Figure 2b, c) are shorter than those of the respective zygotic genes (*C. elegans* [Core: $p<10^{-10}$, Metazoan: $p<10^{-27}$]; *D. rerio* [$p<0.1$, $p<10^{-3}$]) while the genome-wide average is intermediate. Similarly, *G. gallus* all-maternal genes' 5' IGRs are smaller than the genome-wide average (Figure 2e, Core: $p<10^{-5}$, Metazoan: $p<10^{-3}$). Furthermore, *C. elegans* and *D. rerio* maternal and all-maternal gene classes are enriched in core genes compared to the zygotic class ($p<10^{-147}$). This pattern is strikingly reversed in the mammals (Figure 2d, f). Mouse strict-maternal gene 5' IGRs are longer than the genome-wide average (Core: $p<10^{-3}$, Metazoan: $p<10^{-7}$) while the 5' IGR of strict-zygotic genes are smaller (Core: $p<10^{-9}$, Metazoan: $p<0.01$). Similarly, human all-maternal gene 5' IGR lengths are larger than the genome-wide average (Figure 2f, Core: $p<0.03$, Metazoan: $p<10^{-7}$). Unlike the other animals, mouse strict-maternal and all-maternal classes are enriched for metazoan genes ($p<10^{-226}$).

These differences among maternal genes between mammals and the other animals is highlighted by the otherwise consistent relationship observed in all animals of shorter regulatory region lengths for core genes than for metazoan genes (*C. elegans*: $p<10^{-49}$; *D. rerio*: $p<10^{-17}$; *G. gallus*: $p<10^{-29}$; *M. musculus*: $p<10^{-20}$; *H. sapiens*: $p<10^{-5}$). Specifically, as observed in Drosophila, the 3' UTRs of core genes are shorter than the 3'UTRs of metazoan genes and the 3'UTRs of strict-maternal and all-maternal transcripts are underrepresented for short lengths (Figure S2, Figure1 for *G. gallus* and *H. sapiens*). Thus the only significant difference in gene architecture between mammals and the other animals examined here is in length of the 5' IGRs of maternal and zygotic genes. The

relatively large size of mammalian maternal 5'IGRs compared to the genome-wide set suggests that maternal genes in mammals have complex and highly specific transcriptional regulation, whereas maternal genes in the other animals which are much shorter than the genome-wide set are regulated with less specificity.

Mammalian maternal genes are under selective pressure to maintain large 5' IGRs

These observations may reflect either an actual biological difference or a limitation in our definition of maternal and zygotic genes. In all animals, the data for identification of zygotically transcribed genes spanned a time course extending many cell divisions after the start of zygotic transcription, at least up to the metazoan hallmark of gastrulation [4, 9, 15, 33]. It has been suggested that gastrulation and not fertilization, is the time point best fit for alignment of eutherians development with other metazoans [34]. If true, we would expect mouse zygotic genes that are expressed at or after gastrulation to exhibit increased transcriptional complexity. Interestingly, the density of conserved sequences is high in non-coding regions flanking genes expressed in mouse embryos at 9.5-10.5 days of gestation but not earlier in development [25]. Furthermore, genes flanked by gene deserts are enriched in developmental functions in mouse, as well as in human and chicken [26]. This suggests that analysis of IGRs of genes expressed later in mouse development may identify a developmental time point in which the 5' IGR of the genes expressed will be as long, if not longer, than the strict-maternal set. For maternal genes, sparse mRNA abundance measurements may hamper our ability to distinguish strict-maternal-only genes from maternal-zygotic genes.

To confirm that our observations were due to a true biological difference, we compared the all-maternal class from each animal to its respective genome-wide average. For mouse, 5' IGRs of the all-maternal class were larger than the genome-wide average, whereas for all other animals the 5' IGRs of all-maternal genes was statistically significantly shorter than the genome-wide average (Figure S3). These observations highlight that the differences observed in the architecture of maternal genes' 5' IGRs, both when compared to zygotic genes within the same animal and when compared across animals, are due to true biological variation.

The observed differences in gene architecture between mammalian maternal genes and other animals may be due either to the expression of different genes or to differing

regulatory needs of the same genes. Comparative analysis of relative changes in IGRs of maternally expressed versus non-maternally expressed orthologous genes, offers an opportunity to discern the cause of the observed differences. From the animals studied here, *G. gallus* is phylogenetically closest to mammals and unlike them; its maternal genes have short 5' IGRs. To account for differences in absolute genome size we normalized and ranked regulatory region lengths and then calculated the ratio of ranks between individual one-to-one ortholog pairs of chicken-human and mouse-human (see Materials and Methods). For each orthologous pair we obtained one value representing its fold change in percentile ranking of IGR length between chicken and human, and another for its fold change between mouse and human. Comparison of fold changes of all-maternal one-to-one orthologs versus the set of all one-to-one orthologs shows a shift towards larger fold changes in human to chicken (Figure 3, blue lines P-value < 0.01). However, calculating this ratio for mouse versus human genes showed no statistically significant fold changes (Figure 3, red lines). This implies that the 5' IGRs of maternally expressed genes in human and mouse have expanded more than would be expected given the genome sizes or that chicken maternally expressed genes have shrunk. Coupled to the observation that oocyte deposited transcripts in chordates are highly conserved [35], we conclude that the difference in maternal genes' 5' IGR lengths between mammals and other animals may be due to selection for complex transcriptional regulation of mammalian maternal genes.

Discussion

The variations observed across six animals in 5' IGR and 3' UTR lengths provide an opportunity to understand the evolutionary pressures shaping maternal and zygotic genes. To do so, we have relied on the amassed knowledge that precise gene regulation in space, time and abundance, requires complex regulatory regions [36] which in turn requires more genomic real estate [19, 20, 37, 38]. Our observations that in every animal studied here, the regulatory regions of maternal or zygotic core genes are shorter than those of the respective metazoan genes support this notion.

D. melanogaster maternal genes have previously been reported to have significantly longer 3'UTRs than non-maternal genes [5]. However, our meta-analysis of

early embryogenesis in six different species suggests that this statement is inaccurate in a subtle but important manner. Specifically, our analysis suggests that the universal pattern for 3' UTRs of maternal genes is that they are not longer than zygotic genes, but rather for both core and metazoan classes are underrepresented for short lengths. This suggests that the post-transcriptional regulatory constraint imposed on maternally expressed genes has functioned to maintain 3' UTR lengths across the animal kingdom [1-3, 6, 7]. For maternal genes, transcriptional regulatory mechanisms cannot specify spatiotemporal expression patterns therefore any maternal gene that shows complex expression must employ a post-transcriptional regulatory program. Conversely, this regulatory constraint on 3'UTRs maternal genes does not convey any knowledge on the complexity of the regulatory program or require that zygotic genes not utilize post-transcriptional regulatory mechanisms. This is best observed in the De Renzis et al. *D. melanogaster* dataset in which the maternal and zygotic contributions are precisely determined by genetic decoupling (Figure S1). However, it is also apparent in our analysis of *C. elegans* (Figure S2a,b) and *D. rerio* metazoan genes (Figure S2b) , in all of which, the longest 3' UTRs belong to strict-zygotic metazoan genes in agreement with recent works on the role of miRNAs in embryonic development [21, 22, 39].

In contrast, analysis of maternal and zygotic gene 5' IGRs yielded a dichotomy between mammals and the other animals. Given the highly conserved relationship between core and metazoan genes relative 5' IGR regulatory region size, what explains the divide in transcriptional specificity when it comes to maternal genes' transcriptional regulation? An appealing possibility is that differences in gene architecture are mirroring differences in development, specifically pre- and post-fertilization dynamics. We note that the divide in relative 5' IGR size precisely matches the species mode of reproduction. Those with relatively short 5' IGR are all egg laying, oviparous, animals, whereas those with relatively long 5' IGR length are the viviparous mammals. An important difference between oviparous and viviparous animals that is likely to affect gene architecture is the temporal constraint on maternal contributions to the embryo, which for the oviparous species ceases at fertilization, while in the viviparous species continues post-fertilization. To our knowledge, the only other developmental characteristic which corresponds to the differences in regulatory region size is that many

oviparous embryos begin development with a series of rapid cellular cleavages, while in mammals the initial cell cycles are slow, with rapid cleavages occurring only later [34]. Indeed, in animals where initial cleavage division are rapid, early zygotic genes often have small or no-introns [15] a gene architectural feature important for producing a functional transcript during these abbreviated cell cycles [40]. However, the 5' IGR is not transcribed and transcription of the maternal genes occurs before these rapid cleavages, thus the rapid early development can have only an indirect effect on maternal gene architecture.

One mechanism by which developmental constraints, such as rapid early development or a prolonged pre-fertilization stasis, can affect gene architecture is by the selection for or against expression of specific gene classes in either the oocyte or embryo. Wieschaus has proposed that in rapidly developing oviparous animals gene expression is a limiting resource [14]. Under this hypothesis, those genes whose expression can be accommodated from either maternal or zygotic origin, will, over evolutionary timescales, shift to maternal expression. This will relieve the embryo from the synthetic cost (energy and time) to express those genes, thereby minimizing the time to hatching and maximizing the competitive advantage for limited environmental resources. In the extreme, the only transcripts to be expressed zygotically would be those providing spatial and temporal patterning information or whose precocious expression would disrupt early events [14]. The analysis of the high resolution *D. melanogaster* dataset is fully consistent with this hypothesis. Strictly zygotic genes are highly enriched for patterning genes. Similarly, we detect a strong enrichment for metazoan functions, including patterning, in the other oviparous species we analyze. Furthermore, *D. melanogaster* strictly zygotic genes have very large regulatory regions, much larger than the genomic average or even of other developmental genes (Strictly zygotic vs. Developmental genes Core: $p<0.09$, Metazoan $p<10^{-4}$, data not shown). The insight we gain on complex regulation and specificity from the analysis of core and metazoan genes, suggests that the expression of these strictly zygotic genes is temporally and spatially complex. On the other hand, the 5' IGR length (but not 3' UTRs) of maternally expressed genes (including maternal-zygotic and mostly-zygotic) is dramatically shorter than the genomic average,

suggesting reduced regulatory specificity. Again, we observe the same phenomena in the other oviparous species for which zygotic gene data is available (Figure 4a).

Wieschaus hypothesized, for fast-developing oviparous organisms, an efficiency-based shift towards maternal gene expression [14]. However, based on our data we propose that the shift, under certain conditions, can be towards zygotic gene expression. Specifically, viviparous animals develop relatively slowly and the embryo competes for limited environmental resources only via the mother. In contrast, the relatively undifferentiated mammalian oocyte needs to persist indefinitely, and thus may be under selective pressure to minimize energy expenditures and thus maximize gene expression specificity (larger 5' IGRs). Thus, selection for efficiency may generate complex 5' IGRs relative to genome-wide average for viviparous maternal genes and for oviparous zygotic genes.

One of the most striking features of our analysis was the low complexity 5'IGR of maternal genes relative to genome-wide average in oviparous animals. This feature is only partially explained by a shift in functional composition, as it occurs for both core and metazoan gene subclasses as well as in one to one orthologs (Figure 3). We consider two hypotheses to explain this. The first is tolerated profligate expression. The apparently low threshold for maternal expression may enable many genes, over evolutionary time, to non-functionally sample maternal expression. Over time, maternal expression of developmentally neutral genes will accumulate. However, this hypothesis does not explain the apparent selection for non-short 3'UTRs, which suggest selection for post-transcriptional regulatory information. Thus, we propose a second hypothesis: Maternal contributions to embryonic development also include energy and nutrition. Mammals rely on lactation and placentation, while oviparous animals deposit yolk, consisting mainly of proteins, lipids, and phosphorous, into oocytes. The non-functional maternal transcripts provide nutrient stores of nucleotides and phosphate for the rapidly developing embryos. Our data shows a positive correlation between maternally provided nutrition (low for worm and fly*;* higher in zebra fish and chicken; and highest in mammals) and the complexity of maternal gene regulation (Figure 4b). Since maternal transcripts also provide a low osmotic store of nucleotides and phosphate, they may be considered nutritional. Thus, it is possible that some maternal transcripts are purely

nutritional. Such a hypothesis suggests that "misexpressing" a gene in the maternal germline should not be associated with an energy or efficiency cost. Rather, such "profligate" expression of non-detrimental transcripts may be advantageous and selected for. Furthermore, such a selective force could provide a mechanism for creation of new non-coding RNA genes that could evolve into coding genes or exons.

These two interpretations, developmental constraints and nutrient stores, present three testable predictions. First, both models predict a bias in gene function between genes expressed maternally and zygotically. For example, consider a gene that is not selected for either a maternal or a zygotic mode of expression. The expectation is that expression of that gene will drift between strict maternal and strict zygotic expression, such that, at any given time, a set of such genes would be equally represented in both groups. Thus, any bias in the distribution indicates non-neutral evolution, either by functional restriction or gene flow based on energy and timing considerations as described above. Indeed, as we noted above, we observed maternal depletion/zygotic enrichment of metazoan-specific genes, which are enriched for patterning functions, in fast developing embryos (Figure 4b).

Second, the nutrient stores model predicts enrichment for expression of non-functional maternal genes in organisms with limited maternal nutritional contributions (yolk). This is based on the positive correlation we observe between the amount of yolk and the simplicity of maternal gene expression, suggesting that maternal gene regulation becomes promiscuous as maternal nutritional contributions are limited (Figure 4b). Consistent with this, many maternally expressed *C. elegans* and *D. melanogaster* genes do not have an apparent phenotype by RNA interference knockdown [41-43]. In support of this, we tested for regulatory region length differences between *C. elegans* maternal genes for which an RNAi phenotype is detected and those for which it is not (see Materials and Methods). Significant differences were detected in 5' IGR lengths ($p<10^{-8}$), but not 3'UTRs (Figure S4).

Third, we predict that the constituency and regulation of maternal and early zygotic transcripts will only mirror phylogeny to the extent that it agrees with forms of maternal contribution. Viviparity and oviparity have developed multiple independent times, in various forms, in distant branches such as arthropods, sharks, lizards and

eutherian mammals. Based solely on the extent of maternal contribution, our results predict not only how early developmental genes would be regulated in marsupials and monotreme species, relatively close to the studied mammals; but also, that the regulation of genes in early development would be more similar between two distant viviparous animals than between closely related animals with differing reproductive modes.

Conclusions

Here we analyze the regulation constraints of the maternal-zygotic transition, a key developmental process in all animals, involving thousands of genes. The utilization of regulatory region lengths to study complex molecular processes circumvents the present deficiency in detailed information on individual gene regulation and offers a clear methodology, for study of other so-far undecipherable biological processes. Importantly, as a baseline control, we show that differences in the inferred lengths of regulatory regions between different functional gene classes are conserved, irrespective of genome size. At a time when new, non-model organisms' and unannotated genomes are being sequenced at an ever-increasing rate, such methodologies are required to identify and study genes in these organisms.

Our comparative analysis of maternal and zygotic genes within an animal reveals that the location and abundance of regulatory content is driven by at least two forces: one, reflected in the inferred functional complexity of gene action [19, 21] and a second, related to the origin of synthesis of transcripts. This latter selective evolutionary force is acting to modify, as a function of germline versus embryonic transcript synthesis, the gene regulatory architecture of thousands of genes. In contrast, cross-species comparisons, allow analyses of this force and suggest that it is coupled to the timing of the maternal-zygotic transition, which correlates with alternative strategies for managing maternal versus zygotic energy expenditures at the physiological level. Taken together, these results uncover an ancient force affecting the development of all multi-cellular organisms and provide clear predictive criteria for the nature of maternal-zygotic gene regulation in other animals.

Materials and methods

Classification of genes to maternal and zygotic classes

Gene identifiers, chromosomal locations and sequences for all organisms were mined from EnsEMBL V42 December 2006 [44] and Wormbase (Wormbase, Release WS160). To classify genes to either maternal or zygotic origin we used the expression datasets of Baugh et al.[9], De Renzis et al. [15], Giraldez et al. [4] and Hamatani et al. [10] for *C. elegans*, *D. melanogaster, D. rerio* and *M. musculus* respectively. To identify maternal genes in *H. sapiens* and *G. gallus* we used the expression data of Kocabas et al. [12] and Lee et al. [28]. See Additional file 1 for a detailed description how maternal and zygotic genes were identified from each of these datasets.

For *C. elegans*, maternal and zygotic classes correspond respectively to the Strictly Maternal Degrading and Strictly Embryonic classes [9]. For *D. melanogaster*, De Renzis et al. [15] reported, at a fold change of three and a P-value < 0.001, 6,485 genes expressed maternally of which 2,110 decreased significantly in their abundance during the time course. Of the 2,110 genes, 633 had a significant zygotic component contributing to their measured abundance level (Table S7 in De Renzis et al.). We considered the 6,485 genes as all-maternal and the 1477 maternal decreasing genes with no zygotic component as strict-maternal. For the zygotic class, we used the 334 genes expressed at cycle 14 with no maternal contribution (Table S4 in De Renzis et al). The remapping of genes to FlyBase 4.3 reduced the number of genes in each class to 5,923, 1,358 and 314 for all-maternal, strict-maternal and zygotic respectively. For *D. rerio* we used the Giraldez et al. [4] classification of *D. rerio* genes as 'Predominantly Maternal' and 'Predominantly Zygotic' as 'Maternal' and 'Zygotic' classes respectively [4]. Briefly, genes expressed at 1.5 hours post-fertilization and showing a significant reduction at 50% and 90% epiboly were considered maternal. Genes expressed significantly at the 50 and 90% epiboly stages and not at 1.5 hours post-fertilization were considered zygotic. For *G. gallus*, we considered the top ranked 50% of expressed genes at stage X embryos (a laid egg) as maternal. In stage X eggs, an undifferentiated blastoderm has formed on top of the yolk, but major zygotic activation has yet to occur [45]. Results did not change if we set the threshold to a more restrictive 25%, but the number of genes was reduced which affected our orthologous gene comparisons (see

below). For *M. musculus* [10] genes mapping to clusters 7 or 9 were considered 'Maternal' and genes mapping to clusters 1, 4, 5 and 8 as 'Zygotic'. To classify which genes were expressed during gastrulation we ranked genes detected as expressed in wild type embryos from 6.5 days post-cleavage [33]. The top 25% expressed genes were considered zygotic. Varying this threshold from 5% to 50% did not change our results. The 5331 transcripts identified by Kocabas et al. [12] as up regulated in *H. sapiens* MII oocyte transcripts were considered maternal. To the best of our knowledge, a quality dataset identifying human zygotic genes is not available. For each organism the genome-wide gene set was defined as all genes in the genome that meet the criteria (as defined in the Classification of genes to core and metazoan classes and Estimates of regulatory region lengths sections below) to be included in the analysis (e.g. No downstream operon genes were included in the *C. elegans* genome-wide set when calculating the distribution of genome-wide 5' IGR lengths).

Classification of genes to core and metazoan classes

We used the Inparanoid: Eukaryotic Ortholog Groups database (Release 5.1, January 2007) [30] to classify genes into core and metazoan classes by phylogenetic profiling. This version of Inparanoid contains an all against all protein coding gene blast comparison of 26 organisms – 1 prokaryote, 3 unicellular eukaryotes, 2 plants and 20 metazoans (including a urochordate, nematodes, insects, fish, bird, amphibian and mammals) [30]. A core gene was defined as any gene present in one or more of the unicellular organisms included in InParanoid. A metazoan gene was defined as any gene present in 2 or more animals included in Inparanoid that is not present in the core gene set or in plants. The organisms used to define the core gene set are *E. coli, S. cerevisiae, S. pombe* and *D. discoideum.*

We tried several different criteria (higher and lower) for the metazoan gene set definition, and obtained similar qualitative results with different values of significance. For *C. elegans* and *D. melanogaster* we repeated our analysis using the classification scheme defined by Nelson et al. [19] which classifies genes by their expected regulation complexity (simple or complex) based on their molecular functions and the biological processes they are involved in. For *C. elegans* we updated the gene annotations directly from Wormbase GO (Wormbase, Release WS150). For both species, all results obtained

from this analysis were qualitatively the same as the ones obtained from the phylogenetic profiling dataset.

Estimates of regulatory region lengths

We defined a gene's 5' IGR length as the distance between its 5' most coding nucleotide and the closest respective upstream or downstream coding nucleotide belonging to a different gene on either DNA strand. Similarly, 3' IGR length was calculated as the distance from the 3' most stop codon to the downstream closet coding nucleotide belonging to a different gene. We defined the 1st intron is the intron closest downstream to the translation start site. To estimate first intron lengths, we used two measures: the length of the largest first intron of a gene among all the first intron lengths of its alternative splicings, and the largest continuous non-coding segment in the first intron. Both intron length measurement types yielded similar results. In *C. elegans*, for genes transcribed as a part of an operon, only the 5' most gene (first gene) was included in any analysis involving 5' IGR length.

The length of a gene's 3' UTR was approximated as the maximum 3' UTR length of all of its alternatively spliced transcripts. A similar calculation was performed for 5' UTRs. We considered the sum of both 3' and 5' UTRs as the total post-transcriptional regulatory region size for all animals except for *C. elegans*, where post-splicing makes this metric moot. A large fraction of genes in any given genome are annotated with either no UTR information or with a UTR that is only a few base pair long. We noticed that this UTR annotation is replaced with full length UTRs with successive updates of the database and hence appears to be missing or incomplete annotation. No significant enrichment in extremely short UTRs (less than 5 base pairs) was detected for either core, metazoan, maternal or zygotic genes, however their inclusion in the analysis shifted the mean/median of the distributions greatly due to their large numbers. Thus we placed a lower bound on UTR length, considering them as artifacts, and discarded any 3' UTRs below 5 base pairs and any 5' UTRs below 3 base pairs in all species.

We calculated 3' UTR lengths twice, once allowing for multiple exons in the 3' UTR and once without. Roughly 10% of reported 3' UTR in every organism have

multiple 3' UTR exons which are thought to be subject to non-sense mediated decay degradation [46] – statistical tests and plots appearing here are all for 3'UTR which do not contain multiple exons – but results are qualitatively the same when allowing for multiple exons in 3' UTR. For zebra fish, only genes having a RefseqId [47] were included in the analysis of 3' UTR lengths.

To determine that our results are robust to exact definitions of regulatory region lengths, we considered for both transcriptional and post-transcriptional regions alertnative definitions of a genes' regulatory regions. For transcriptional regulatory region lengths comparisons between gene groups, we performed our analysis using not only 5' IGRs, but also the total length of a gene's 5' IGR + the 1st intron, the sum of IGRs (5' IGR + 3' IGR), and the sum of all three (5' IGR + 1st intron + 3'IGR). For post-transcriptional regulation we estimated the 3' UTR length as well as the total sum of UTR (5' + 3'). Transcriptional regulatory region estimates across all genes showed that they were highly correlated with one another (Figure S5a). Similarly, the two post-transcriptional regulatory region estimates were also highly correlated with one another (Figure S5b). We applied the analyses presented here using each of the different estimates of transcriptional and post-transcriptional regulatory region length for each of the species. Analysis of each of these for every species yielded qualitatively the same results. The 5' IGR + the 1st intron analysis mirrored very closely the observed signal in the 5' IGR, whereas analysis of regions which included the 3' IGR showed reduced, but still significant, differences between regions. Similarly, considering the sum of the 5' UTR and 3' UTR regions for post-transcriptional regulation yielded similar results qualitatively and significance wise. Thus, the results of the analyses we present are robust to the exact definition of regulatory region length, at least to a degree matching the present knowledge of the location of a gene's regulatory information.

Differences in regulatory region lengths between gene classes

Differences in distributions of the different maternal and zygotic classes were quantified using the non-parametric, one-sided two-sample Kolmogorov-Smirnov test at a significance level $\alpha = 0.05$. The Kolmogorov-Smirnov test tests the null hypothesis that

two sample distributions are drawn from the same distribution and does so by quantifying the distance between the two empirical cumulative distributions. For a given comparison of two distributions, the reported significant p-values for this one-sided test indicate that the 1st distribution of regulatory region lengths under evaluation is shifted to the right (i.e. fewer shorter lengths) of the 2nd. For both transcriptional and post-transcriptional regulatory regions, we performed this test once when considering all (100%) regulatory region lengths within each group.

In addition, to quantify the extent that maternal UTRs are under-represented for short lengths, we iteratively applied a one-sided two-sample Kolmogorov-Smirnov test on defined subsets of the distributions. The subsets were determined empirically beginning with the 15th percentile and incrementing by 5%. For each comparison we report the top most percentile that produced a p-value < 0.05 and we identify the percentile at which the minimum (most significant) p-value was detected. In the text we report only the top-most significant percentile, whereas in the figure legend we report the most significant p-value and the accompanying percentile as well as the top-most significant percentile.

Orthologous gene analysis between chicken and mammals

To account for differences in genome size and gene number, within one genome, we rank-ordered and normalized all genes by 5' IGR length. We then identified all genes with single orthologs in human, mouse and chicken (1:1:1) using Inparanoid. For these, we calculated the ratio of 5' IGR length ranks between every human gene and its one-to-one orthologous chicken counterpart. This ratio represents the fold change in percentile ranking. This procedure was repeated for human and mouse genes. For both chicken-human and mouse-human, these were then divided into those orthologs classified as all-maternal in both species and the remaining orthologous genes. Thus, for every gene we obtained one value representing its fold change in percentile ranking between chicken and human, and another for its fold change between mouse and human.

Developmental constraints and nutritional/promiscuity model analysis

To perform this analysis, animals were placed into one of three classes (small, medium, large), based on the estimated nutritional contribution provided by the mother. This was estimated from the ratio of the size of an oocyte to the size of an embryo at gastrulation. For each animal the extent of maternal gene transcriptional regulatory complexity is estimated by the ratio of maternal metazoan gene 5' IGR length to the genome average. We restricted our comparison to metazoan genes, as they are the subset most reflective of changes in regulatory complexity. To calculate the ratio of maternal to genome-wide regulatory region lengths for strict-maternal genes, we used three different measures, including the median of each gene class, the 75 percentile and a 5% trimmed mean. For *G. gallus* and *H. sapiens* we used the all maternally expressed genes in substitute for a strict-maternal class, which has not been defined.

To test for differences in regulatory region length between *C. elegans* maternal genes for which an RNAi phenotype is detected and those for which it is not, we obtained from Wormbase WS200 a list of 700,0000 *C. elegans* RNAi tests of function each of which was annotated as to whether a phenotype was observed or not. Cross-checking this against the maternal gene expression list yielded 114,789 that were performed on one of the 5,591 all-maternally expressed genes. Classifying these genes by whether or not a phenotype was observed for them in an RNAi experiment yielded 922 genes which showed no observed phenotype (presumed non-functional maternal genes) and 4669 with one or more functional (with phenotype) maternally expressed genes. Regulatory region lengths comparisons between the two groups were performed as detailed in the "Differences in regulatory region lengths between gene classes" subsection of Materials and Methods.

Abbreviations

IGR – Intergenic region; UTR – Untranslated Region

Authors' contributions

SSO conceived, designed and performed the study, analyzed the results and drafted the manuscript. YP participated in the study design, analysis of the results and drafting of the

manuscript, CPH conceived and designed the study, analyzed the results and drafted the manuscript. All authors read and approved the final manuscript.

Acknowledgements

We thank A. Jose, R. Milo, R. Baugh, I. Yanai and J. Whangbo for comments on the manuscript and helpful discussions. M. Kirschner. D. Haig, D. Petrov, G. Bejerano and A. Giraldez for helpful ideas. J. Cuff and M. Ethier for IT support. This work was supported in part by NIH grant GM064429 to CPH.

References

1. Pique M, Lopez JM, Foissac S, Guigo R, Mendez R: **A combinatorial code for CPE-mediated translational control.** *Cell* 2008, **132:**434-448.
2. Gavis ER, Lehmann R: **Translational regulation of nanos by RNA localization.** *Nature* 1994, **369:**315-318.
3. Merritt C, Rasoloson D, Ko D, Seydoux G: **3' UTRs Are the Primary Regulators of Gene Expression in the C. elegans Germline.** *Curr Biol* 2008, **18:**1476-82.
4. Giraldez AJ, Mishima Y, Rihel J, Grocock RJ, Van Dongen S, Inoue K, Enright AJ, Schier AF: **Zebrafish MiR-430 promotes deadenylation and clearance of maternal mRNAs.** *Science* 2006, **312:**75-79.
5. Tadros W, Goldman AL, Babak T, Menzies F, Vardy L, Orr-Weaver T, Hughes TR, Westwood JT, Smibert CA, Lipshitz HD: **SMAUG is a major regulator of maternal mRNA destabilization in Drosophila and its translation is activated by the PAN GU kinase.** *Dev Cell* 2007, **12:**143-155.
6. Seydoux G, Fire A: **Soma-germline asymmetry in the distributions of embryonic RNAs in Caenorhabditis elegans.** *Development* 1994, **120:**2823-2834.
7. Schier AF: **The maternal-zygotic transition: death and birth of RNAs.** *Science* 2007, **316:**406-407.

8. Arbeitman MN, Furlong EE, Imam F, Johnson E, Null BH, Baker BS, Krasnow MA, Scott MP, Davis RW, White KP: **Gene expression during the life cycle of Drosophila melanogaster.** *Science* 2002, **297:**2270-2275.
9. Baugh LR, Hill AA, Slonim DK, Brown EL, Hunter CP: **Composition and dynamics of the Caenorhabditis elegans early embryonic transcriptome.** *Development* 2003, **130:**889-900.
10. Hamatani T, Carter MG, Sharov AA, Ko MS: **Dynamics of global gene expression changes during mouse preimplantation development.** *Dev Cell* 2004, **6:**117-131.
11. Mathavan S, Lee SG, Mak A, Miller LD, Murthy KR, Govindarajan KR, Tong Y, Wu YL, Lam SH, Yang H, Ruan Y, Korzh V, Gong Z, Liu ET, Lufkin T: **Transcriptome analysis of zebrafish embryogenesis using microarrays.** *PLoS Genet* 2005, **1:**260-276.
12. Kocabas AM, Crosby J, Ross PJ, Otu HH, Beyhan Z, Can H, Tam WL, Rosa GJ, Halgren RG, Lim B, Fernandez E, Cibelli JB: **The transcriptome of human oocytes.** *Proc Natl Acad Sci U S A* 2006, **103:**14027-14032.
13. Azumi K, Sabau SV, Fujie M, Usami T, Koyanagi R, Kawashima T, Fujiwara S, Ogasawara M, Satake M, Nonaka M, Wang HG, Satou Y, Satoh N: **Gene expression profile during the life cycle of the urochordate Ciona intestinalis.** *Dev Biol* 2007, **308:**572-582.
14. Wieschaus E: **Embryonic transcription and the control of developmental pathways.** *Genetics* 1996, **142:**5-10.
15. De Renzis S, Elemento O, Tavazoie S, Wieschaus EF: **Unmasking activation of the zygotic genome using chromosomal deletions in the Drosophila embryo.** *PLoS Biol* 2007, **5:**e117.
16. Yuh CH, Bolouri H, Davidson EH: **Genomic cis-regulatory logic: experimental and computational analysis of a sea urchin gene.** *Science* 1998, **279:**1896-1902.
17. Johnstone O, Lasko P: **Translational regulation and RNA localization in Drosophila oocytes and embryos.** *Annu Rev Genet* 2001, **35:**365-406.

18. Birney E, Stamatoyannopoulos JA, Dutta A, Guigo R, Gingeras TR, Margulies EH, Weng Z, Snyder M, Dermitzakis ET, Thurman RE, Kuehn MS, Taylor CM, Neph S, Koch CM, Asthana S, Malhotra A, Adzhubei I, Greenbaum JA, Andrews RM, Flicek P, Boyle PJ, Cao H, Carter NP, Clelland GK, Davis S, Day N, Dhami P, Dillon SC, Dorschner MO, Fiegler H, et al.: **Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project.** *Nature* 2007, **447:**799-816.

19. Nelson CE, Hersh BM, Carroll SB: **The regulatory content of intergenic DNA shapes genome architecture.** *Genome Biol* 2004, **5:**R25.

20. Walther D, Brunnemann R, Selbig J: **The Regulatory Code for Transcriptional Response Diversity and Its Relation to Genome Structural Properties in A. thaliana.** *PLoS Genet* 2007, **3:**e11.

21. Stark A, Brennecke J, Bushati N, Russell RB, Cohen SM: **Animal MicroRNAs confer robustness to gene expression and have a significant impact on 3'UTR evolution.** *Cell* 2005, **123:**1133-1146.

22. Shalgi R, Lieber D, Oren M, Pilpel Y: **Global and local architecture of the mammalian microRNA-transcription factor regulatory network.** *PLoS Comput Biol* 2007, **3:**e131.

23. Eisenberg E, Levanon EY: **Human housekeeping genes are compact.** *Trends Genet* 2003, **19:**362-365.

24. Vinogradov AE: **Compactness of human housekeeping genes: selection for economy or genomic design?** *Trends Genet* 2004, **20:**248-253.

25. Sironi M, Menozzi G, Comi GP, Cagliani R, Bresolin N, Pozzoli U: **Analysis of intronic conserved elements indicates that functional complexity might represent a major source of negative selection on non-coding sequences.** *Hum Mol Genet* 2005, **14:**2533-2546.

26. Ovcharenko I, Loots GG, Nobrega MA, Hardison RC, Miller W, Stubbs L: **Evolution and functional classification of vertebrate gene deserts.** *Genome Res* 2005, **15:**137-145.

27. Vinogradov AE: **"Genome design" model: evidence from conserved intronic sequence in human-mouse comparison.** *Genome Res* 2006, **16:**347-354.

28. Lee BR, Kim H, Park TS, Moon S, Cho S, Park T, Lim JM, Han JY: **A set of stage-specific gene transcripts identified in EK stage X and HH stage 3 chick embryos.** *BMC Dev Biol* 2007, **7:**60.

29. Sironi M, Menozzi G, Comi GP, Cereda M, Cagliani R, Bresolin N, Pozzoli U: **Gene function and expression level influence the insertion/fixation dynamics of distinct transposon families in mammalian introns.** *Genome Biol* 2006, **7:**R120.

30. O'Brien KP, Remm M, Sonnhammer EL: **Inparanoid: a comprehensive database of eukaryotic orthologs.** *Nucleic Acids Res* 2005, **33:**D476-480.

31. Richter JD: **Cytoplasmic polyadenylation in development and beyond.** *Microbiol Mol Biol Rev* 1999, **63:**446-456.

32. Tadros W, Westwood JT, Lipshitz HD: **The mother-to-child transition.** *Dev Cell* 2007, **12:**847-849.

33. Morkel M, Huelsken J, Wakamiya M, Ding J, van de Wetering M, Clevers H, Taketo MM, Behringer RR, Shen MM, Birchmeier W: **Beta-catenin regulates Cripto- and Wnt3-dependent gene expression programs in mouse axis and mesoderm formation.** *Development* 2003, **130:**6283-6294.

34. O'Farrell PH, Stumpff J, Su TT: **Embryonic cleavage cycles: how is a mouse like a fly?** *Curr Biol* 2004, **14:**R35-45.

35. Evsikov AV, Graber JH, Brockman JM, Hampl A, Holbrook AE, Singh P, Eppig JJ, Solter D, Knowles BB: **Cracking the egg: molecular dynamics and evolutionary aspects of the transition from the fully grown oocyte to embryo.** *Genes Dev* 2006, **20:**2713-2727.

36. Davidson EH: *Genomic Regulatory Systems. Development and Evolution.* San Diego, CA: Academic Press; 2001.

37. Lee S, Kohane I, Kasif S: **Genes involved in complex adaptive processes tend to have highly conserved upstream regions in mammalian genomes.** *BMC Genomics* 2005, **6:**168.

38. Vinogradov AE: **'Genome design' model and multicellular complexity: golden middle.** *Nucleic Acids Res* 2006, **34:**5906-5914.

39. Farh KK, Grimson A, Jan C, Lewis BP, Johnston WK, Lim LP, Burge CB, Bartel DP: **The widespread impact of mammalian MicroRNAs on mRNA repression and evolution.** *Science* 2005, **310:**1817-1821.

40. Rothe M, Pehl M, Taubert H, Jackle H: **Loss of gene function through rapid mitotic cycles in the Drosophila embryo.** *Nature* 1992, **359:**156-159.

41. Kamath RS, Fraser AG, Dong Y, Poulin G, Durbin R, Gotta M, Kanapin A, Le Bot N, Moreno S, Sohrmann M, Welchman DP, Zipperlen P, Ahringer J: **Systematic functional analysis of the Caenorhabditis elegans genome using RNAi.** *Nature* 2003, **421:**231-237.

42. Perrimon N, Engstrom L, Mahowald AP: **Zygotic lethals with specific maternal effect phenotypes in Drosophila melanogaster. I. Loci on the X chromosome.** *Genetics* 1989, **121:**333-352.

43. Perrimon N, Lanjuin A, Arnold C, Noll E: **Zygotic lethal mutations with maternal effect phenotypes in Drosophila melanogaster. II. Loci on the second and third chromosomes identified by P-element-induced mutations.** *Genetics* 1996, **144:**1681-1692.

44. Birney E, Andrews D, Caccamo M, Chen Y, Clarke L, Coates G, Cox T, Cunningham F, Curwen V, Cutts T, Down T, Durbin R, Fernandez-Suarez XM, Flicek P, Graf S, Hammond M, Herrero J, Howe K, Iyer V, Jekosch K, Kahari A, Kasprzyk A, Keefe D, Kokocinski F, Kulesha E, London D, Longden I, Melsopp C, Meidl P, Overduin B, et al.: **Ensembl 2006.** *Nucleic Acids Res* 2006, **34:**D556-561.

45. Zagris N, Matthopoulos D: **Stage-specific gene expression in early chick embryo.** *Dev Genet* 1989, **10:**333-338.

46. Lewis BP, Green RE, Brenner SE: **Evidence for the widespread coupling of alternative splicing and nonsense-mediated mRNA decay in humans.** *Proc Natl Acad Sci U S A* 2003, **100:**189-192.

47. Pruitt KD, Tatusova T, Maglott DR: **NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins.** *Nucleic Acids Res* 2005, **33:**D501-504.

Figure legends

Figure 1 - 3' UTRs of maternal genes are under represented for short lengths.

3' UTR lengths in six animals comparing all maternally expressed (A) core or (B) metazoan genes (solid) versus all other core or metazoan genes in the genome (dotted). (A) Core genes. *C. elegans:* [minimum $p<10^{-18}$, percentile at which the minimum p-value was detected: 20^{th}, , top most percentile showing significance: 100%], *D. melanogaster*: [$p<10^{-9}$,25,100], *D. rerio*: [$p<10^{-6}$,20,85], *G. gallus*: [$p<10^{-5}$,65,100], *M. musculus*: [$p<10^{-12,}$25,100], *H. sapiens*: [$p<10^{-12}$,25,100]. (B) Metazoan genes. *C. elegans*: [$p<10^{-26,}$20,100], *D. melanogaster*: [$p<10^{-30}$,35,100], *D. rerio*: [$p<10^{-6}$,45,100], *G. gallus*: [$p<10^{-17}$,40,100], *M. musculus* [$p<10^{-23}$,20,100], *H.* sapiens: [$p<10^{-18}$,35,100].

Figure 2 - 5' IGR length in all animals is dependent on both gene functional complexity and transcript origin of synthesis.

(A) Genetic manipulation of *D. Melanogaster* enables quantification of maternal and zygotic component of mRNA abundance, allowing analysis of five gene classes. Genes expressed solely by the zygote have long 5' IGRs, whereas genes expressed by the mother have short 5' IGRs. Observed differences are greatest when comparing genes expressed exclusively from one origin. (B-D) Similar comparisons of *C. elegans*, *D. rerio* and *M. musculus* where gene classification is based solely on characteristic strict maternal and strict-zygotic expression profiles. In mouse an inverse relationship between maternal and zygotic genes is observed. (E,F) 5' IGR length comparison of all maternally expressed genes in *G. Gallus* and *H. sapiens* to all other genes in the genome. Like mouse, human maternal genes have large 5' IGRs. In all plots, genes are partitioned to core and metazoan classes by phylogenetic filtering. Core genes have shorter 5' IGRs than metazoan ones. Numbers in parentheses to the right of each box plot bar are the number of genes per class.

Figure 3 - Systematic change in relative size of 5' IGRs of maternally expressed human and chicken one-to-one orthologs.

Shown is the cumulative distribution of fold-change difference in relative 5' IGR size for all human, chicken and mouse 1:1:1 orthologs (dotted) versus those expressed maternally in all three organisms (solid). Fold change is shown on a $\log_2$ axis. A fold change of zero

implies that the length of the 5' IGRs of a gene and its 1:1 ortholog ranked the same within their respective genome. Similarly, a positive fold change implies a gene's 5' IGR has either expanded in relative size in human (and/or shrunk in mouse or chicken) with respect to the relative size of its ortholog's 5' IGR in mouse or chicken. The converse is implied by negative $\log_2$(fold change).

Figure 4 - Specificity of expression of maternally expressed genes correlates positively with the amount of maternal nutritional contribution.

(A) Schematic summarizing the size of transcriptional regulatory regions of maternal and zygotic genes in each species, relative to one another and to the genome-wide average. We note a dichotomy that matches the reproductive mode. The highly conserved relationship between core and metazoan genes relative 5' IGR regulatory region size suggests that regulatory region length may be considered as a metric for complexity and specificity of transcriptional regulation. (B) Organizing animals by the amount nutritional contribution provided by the mother (small, medium, large) we estimate the specificity of maternal gene expression by the ratio of maternal metazoan gene 5' IGR length to the genome average. Shown are three measures of the ratio of maternal to genome-wide regulatory region lengths for strict-maternal genes (for *G. Gallus* and *H. sapiens* all maternally expressed genes). Comparison is restricted to metazoan genes, as they are the subset most reflective of changes in regulatory complexity.

Additional files

Additional file 1 – Additional Documentation and Figures

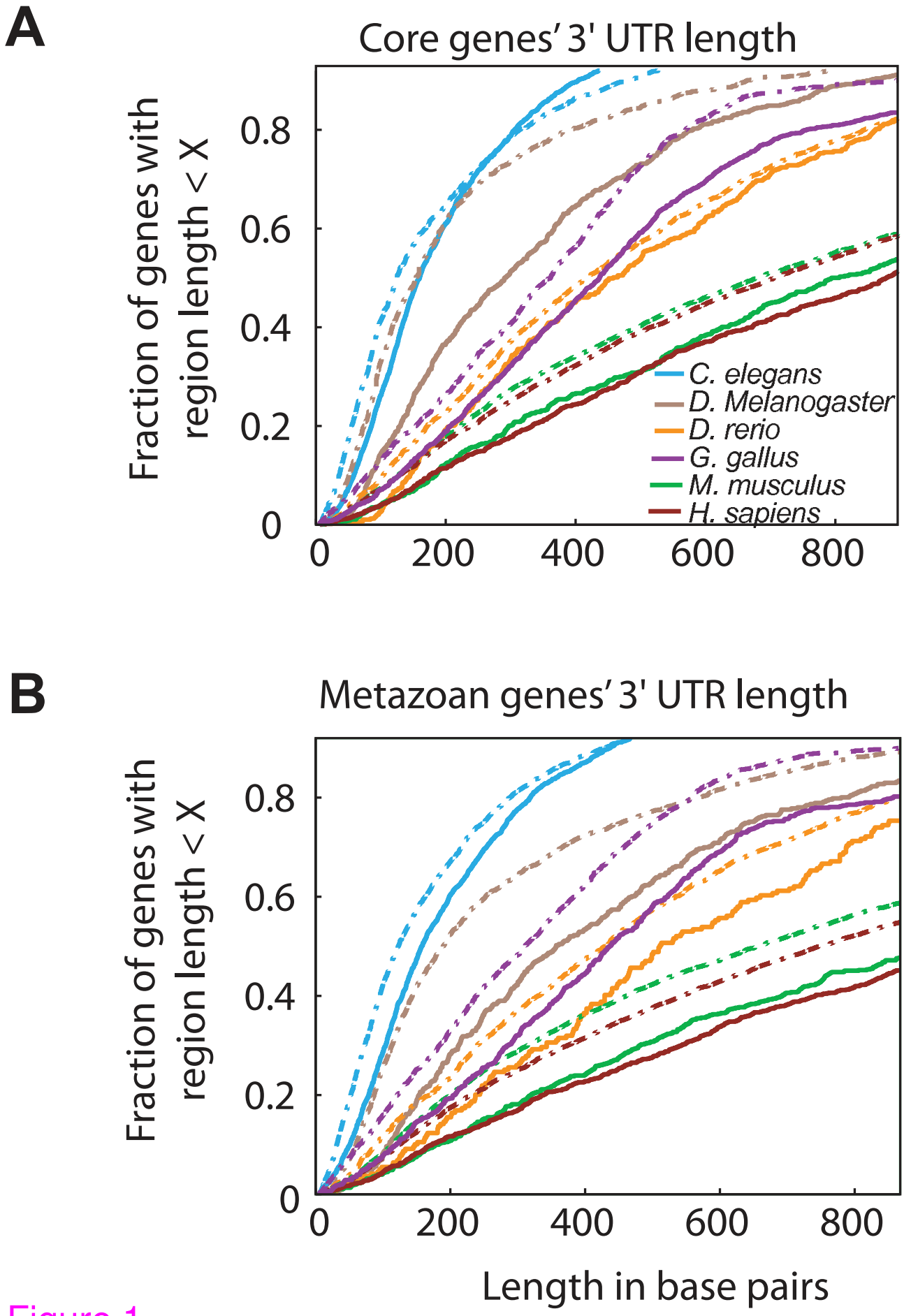

A
Core genes' 3' UTR length
Fraction of genes with region length < X
0
0.2
0.4
0.6
0.8
0
200
400
600
800
C. elegans
D. Melanogaster
D. rerio
G. gallus
M. musculus
H. sapiens
B
Metazoan genes' 3' UTR length
Fraction of genes with region length < X
0
0.2
0.4
0.6
0.8
0
200
400
600
800
Length in base pairs

Figure 1

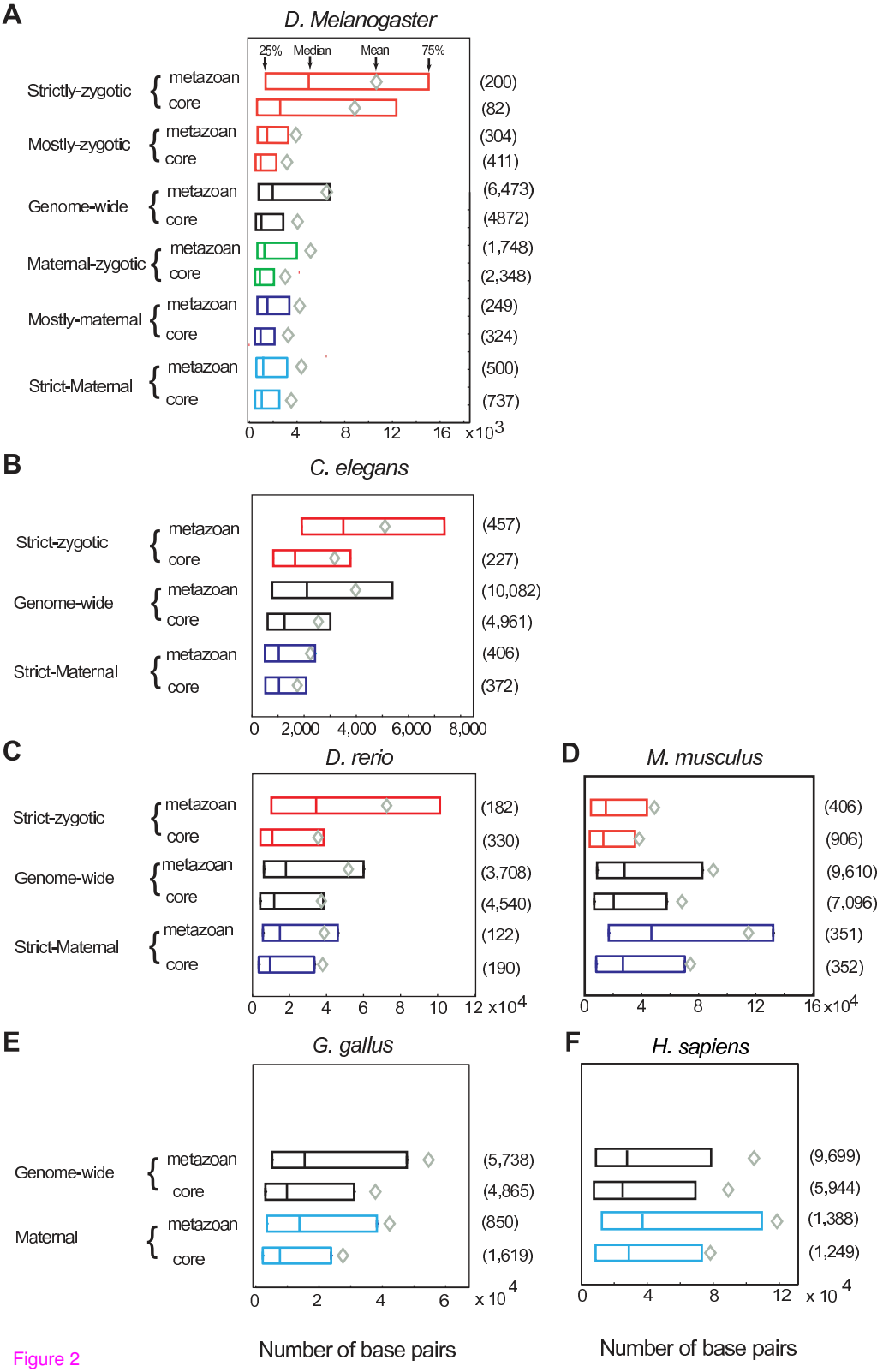

A
D. Melanogaster
25%
Median
Mean
75%
Strictly-zygotic
metazoan
core
(200)
(82)
Mostly-zygotic
metazoan
core
(304)
(411)
Genome-wide
metazoan
core
(6,473)
(4872)
Maternal-zygotic
metazoan
core
(1,748)
(2,348)
Mostly-maternal
metazoan
core
(249)
(324)
Strict-Maternal
metazoan
core
(500)
(737)
0 4 8 12 16 x10^3
B
C. elegans
Strict-zygotic
metazoan
core
(457)
(227)
Genome-wide
metazoan
core
(10,082)
(4,961)
Strict-Maternal
metazoan
core
(406)
(372)
0 2,000 4,000 6,000 8,000
C
D. rerio
Strict-zygotic
metazoan
core
(182)
(330)
Genome-wide
metazoan
core
(3,708)
(4,540)
Strict-Maternal
metazoan
core
(122)
(190)
0 2 4 6 8 10 12 x10^4
D
M. musculus
(406)
(906)
(9,610)
(7,096)
(351)
(352)
0 4 8 12 16 x10^4
E
G. gallus
Genome-wide
metazoan
core
(5,738)
(4,865)
Maternal
metazoan
core
(850)
(1,619)
0 2 4 6 x 10^4
Number of base pairs
F
H. sapiens
(9,699)
(5,944)
(1,388)
(1,249)
0 4 8 12 x 10^4
Number of base pairs

Figure 2

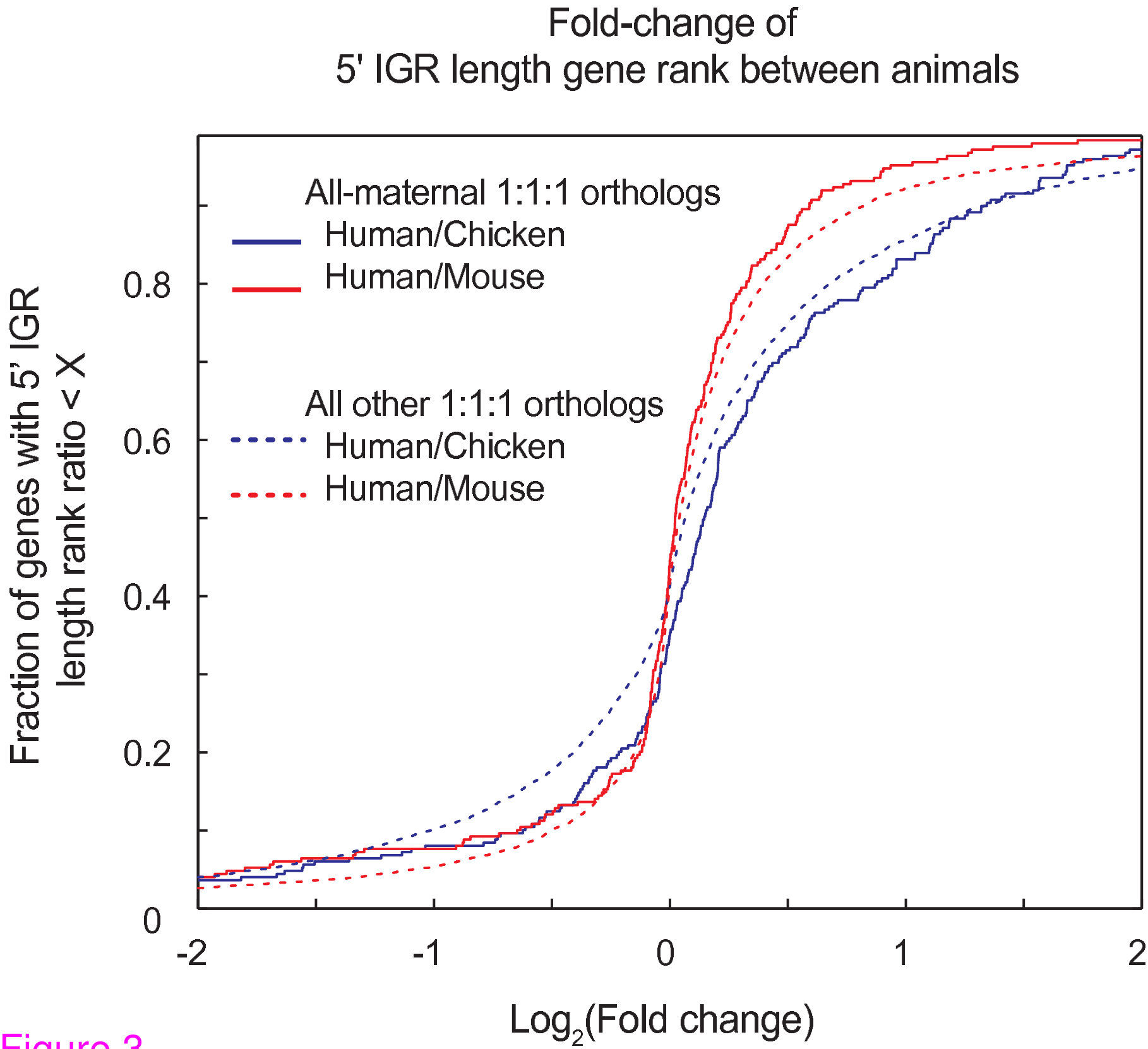

Fold-change of
5' IGR length gene rank between animals
All-maternal 1:1:1 orthologs
Human/Chicken
Human/Mouse
All other 1:1:1 orthologs
Human/Chicken
Human/Mouse
Fraction of genes with 5' IGR length rank ratio < X
0
0.2
0.4
0.6
0.8
-2
-1
0
1
2
Log2(Fold change)

Figure 3

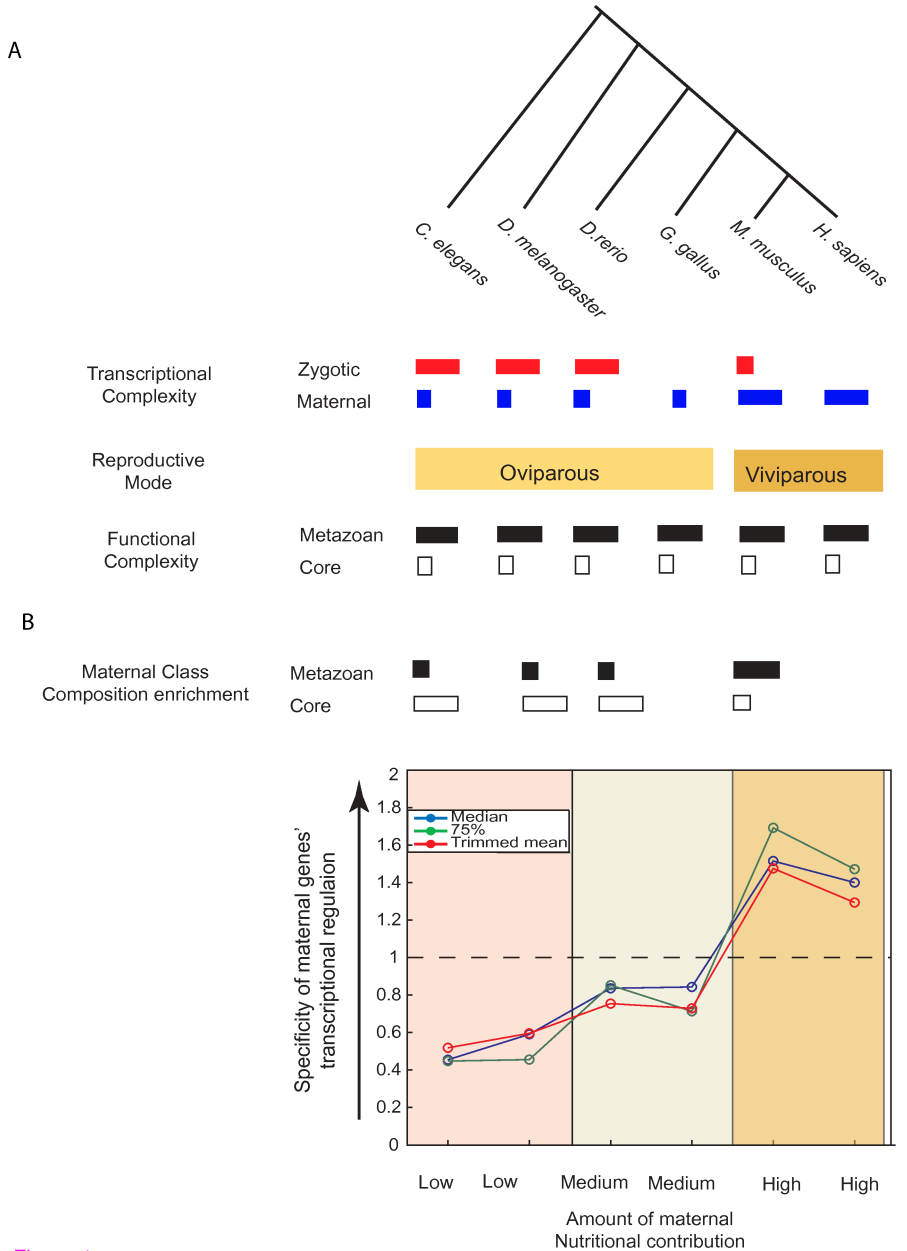

A
C. elegans
D. melanogaster
D.rerio
G. gallus
M. musculus
H. sapiens
Transcriptional Complexity
Zygotic
Maternal
Reproductive Mode
Oviparous
Viviparous
Functional Complexity
Metazoan
Core
B
Maternal Class Composition enrichment
Metazoan
Core
Median
75%
Trimmed mean
Specificity of maternal genes' transcriptional regulaion
2
1.8
1.6
1.4
1.2
1
0.8
0.6
0.4
0.2
0
Low
Low
Medium
Medium
High
High
Amount of maternal Nutritional contribution

Figure 4

Additional Documentation: Composition and Regulation of Maternal and Zygotic Transcriptomes Reflects Species Specific Reproductive Mode

Shai S. Shen-Orr[1†], Yitzhak Pilpel[2], Craig P. Hunter[1§]

[1] Department of Molecular and Cellular Biology, Harvard University, 16 Divinity Ave., Cambridge, MA, 02138, U.S.A

[2]Department of Molecular Genetics, Weizmann Institute of Science, Rehovot, 76100, Israel

[†] Present Address: Departments of Pediatrics and Microbiology & Immunology, Stanford University, Stanford, CA 94305, USA

[§]Corresponding author

This additional documentation contains:

1) Additional figures and legends
2) Detailed description of the choice of public datasets and genes used for the analysis described in the main text.

Additional figure legends

Figure S1 - Comparison of 3' UTR lengths for five maternal and zygotic classes of *D. melanogaster* genes. The 3' UTRs of maternal and zygotic class genes are significantly longer than those of other genes in the genome. Both core and metazoan strict-maternal genes are underrepresented for short 3' UTRs compared to all other classes (Strict-maternal versus Mostly-maternal [Core: $p<0.02$, Metazoan: $p<10^{-6}$]; maternal-zygotic [$p<10^{-24}$, $p<10^{-8}$]; mostly-zygotic [$p<10^{-24}$, $p<10^{-8}$]; strictly zygotic [Not Signifcant, $p<10^{-5}$]; and all coding genes [$p<10^{-30}$, $p<10^{-43}$]), yet the longest 3' UTRs are those of zygotic genes. (a) Core and (b) Metazoan genes.

Figure S2 - Comparison of maternal vs. zygotic 3' UTR lengths for *C. elegans*, *D. rerio*, and *M. musculus* genes. 3'UTRs of maternal genes are underrepresented for short lengths; however, in all animals but mouse, the longest 3' UTRs are zygotic. (a) Core genes in *C. elegans:* [Vs. Genome-wide $p< 10^{-15}$,30%, 100% ; Vs. Zygotic $p<10^{-14}$,40%, 100%], *D. rerio*: [$p< 10^{-4}$,20%, 70% ; Not significant] and *M. musculus*: [$p< 10^{-12}$,40%, 100% ; $p<10^{-17}$,45%,100%], (b) Metazoan genes in *C. elegans:* [$p< 10^{-9}$,25%, 100% ; Not significant], *D. rerio*: [$p< 10^{-6}$,45%, 100% ;Not significant] and *M. musculus*: [$p< 10^{-19}$,25%, 100% ; $p<10^{-9}$,55%,100%].

Figure S3 - Genes expressed in mammalian oocytes have large 5' IGRs. Comparison of 5' IGR lengths of all-maternal core and metazoan genes to the genome wide average. The 5' IGRs of maternally expressed genes in mouse are larger than the genome average (Core: $p< 0.05$, Metazoan: $p<10^{-4}$). This is the opposite of the relationship observed in *C. elegans, D. melanogaster*, and *D. rerio,* (see Figure 2) where the 5'IGRs are smaller than the genome wide average (*C. elegans* [Core: $p<10^{-15}$, Metazoan: $p<10^{-17}$], *D. melanogaster* [$p<10^{-6}$, $p<10^{-17}$], *D.*

rerio [$p<10^{-3}$, not significant for metazoan genes]). This analysis illustrates that the phenomena of large 5' IGRs in mammals is independent of the definition zygotic activation of transcription.

Figure S4 – Maternal genes with no RNAi phenotype have significantly smaller 5' IGR. *C. elegans* all-maternal genes were segregated into two groups based on whether or not a phenotype was observed when they were knocked-down by RNAi. This yielded 922 non-functional (no phenotype) and 4669 functional (with phenotype) maternally expressed genes. In agreement with the nutritional/developmental constraints model, (A) maternal genes with no discernable RNAi phenotype showed significant smaller 5' IGR lengths ($p<10^{-8}$). Yet, (B) no significant differences in 3' UTR lengths were detected, suggesting that all maternal genes, irrespective of function, are regulated post-transcriptionally.

Figure S5 - Different regulatory regions length metrics are highly correlated in all animals. All *D. melanogaster* core and metazoan genes are sorted by their (A) 5' IGR length (top panel). Lengths were normalized to between 0 and 1 by the gene with the longest 5' IGR. Shown below are normalized alternative regulatory region sizes we considered as a proxy for transcriptional regulation complexity, sorted by 5' IGR lengths. These include: the sum of the 5' IGR and 1st intron, the sum of the 5' IGR and 3' IGR or the sum of the 5' and 3' IGRs as well as the 1st intron. ρ denotes the Spearman's rank correlation between the 5' IGR and each of the alternative transcriptional regulatory region size metrics. (B) Same as (A) above only for post-transcriptional regulatory region complexity, for which we considered 3' UTRs (top panel) and the sum of the 5' and 3' UTRs. Similarly high correlations to the ones shown here were observed in all other studied species: *C. elegans*. [5' IGR + 1st intron ρ=.96, 5' +3' IGR ρ=.80, 5'IGR + 1st intron + 3' IGR ρ=.79; 5'+3' UTR Not applicable], *D. rerio* [ρ=.96, ρ=.72, ρ=.71; ρ=.95], *G. gallus* [ρ=.96, ρ=.81, ρ=.80; ρ=.92], *H. sapiens* [ρ=.96, ρ=.82, ρ=.81; ρ=.90].

Detailed description of the datasets and genes used for the analysis described in the main text

All EnsEMBL data was mined using the EnsEMBL perl API [44]. Operon classification for *C. elegans* was obtained from Wormbase (Wormbase, Release WS160). For the worm data set [9], we remapped the custom designed Affymetrix probes to Wormbase gene annotation (Wormbase, Release WS160) and considered only probes which map to single genes. If multiple probes exist per gene, we only used the data obtained from the single gene probe mapping to the 3' most part of said gene. Thus of the reported 8890 reproducibly detected probes we used only 8080 probes each with a 1:1 mapping to a gene.

The *D. melanogaster* dataset [15] was obtained using the Affymetrix Drosophila genome I array which covers 13,500 genes based on Flybase version 1 models. Many gene models have changed since that version 1 of Flybase and so we restricted ourselves to analyzing genes which we could map to FBIds of Flybase, Release 4.3 and having a CG model number. We also performed all of the presented analysis using the embryonic portion of the Arbeitman et al. [8] dataset. All results obtained were qualitatively the same but usually with a lower significance due to the small number of genes covered by the array. In this case we defined zygotic genes as all those expressed during embryogenesis but not detected maternally. Our choice to present the analysis using the De Renzis et al. dataset is due to its higher coverage of the genome and the ability to separate zygotic and maternal contributions to transcript abundance level.

The *D. rerio* array used to detect expressed genes covers less than half of all predicted genes [4]. Only genes present in contigs that have been mapped to chromosomes were included. We performed an analysis comparing zebrafish maternal and zygotic genes using a different dataset

published by Mathavan et al.[11]. This yielded qualitatively similar results, with a higher significance of differences between 5' IGR of maternal and zygotic genes, but a lower significance in 3' UTRs. Since the results published by Mathavan et al. were obtained using a two-channel cDNA array with pooled RNA from all developmental stages as reference, we elected to use the Giraldez et al. dataset instead.

The preimplantation *Mus musculus* gene expression data [10] was generated on the custom NIA 22K 60-mer Oligo Microarray. We used Entrez gene identifiers to map microarray probes to specific mouse EnsEMBL genes, and did not consider probes for which no mapping was found. Probes mapping to multiple genes were not considered. For genes that mapped to multiple probes, a single probe was selected for maternal/zygotic classification, if the expression profile of all probes mapped to the same mega-clusters (See section on Classification of genes to maternal and zygotic classes in Materials and Methods of the main text). This reduced the probe number to 7187 each of which maps to a single gene covering 30% of all protein coding genes in the mouse genome. However, the NIA 22K 60-mer Oligo microarray is enriched for genes expressed in stem cells and preimplantation embryos [10] so we expect that our results are based on a higher proportion of genes expressed in the oocyte than the 30% of the covered genome. To detect which genes are expressed during mouse gastrulation we used wild type samples from dpc 6.5 of mouse development generated as controls in a study conducted by Morkel et al. [33]. The array used in this work is the Affymetrix murine 11K which has been designed based on Unigene Build 4. Mapping Affymetrix probeset names to NCBI assembly 36 of the mouse genome identifies 6040 genes covering roughly 24% of mouse coding genes. We eliminated probes not matching EnsEMBL V45 genes as well as probes mapping to more than one gene and averaged

probes mapping to the same gene. Finally we calculated the mean expression value for every gene across the three hybridizations.

Kocabas et al.[12] profiled human oocytes of young reproductively healthy females using the Affymetrix U133 Plus 2.0 Genechip, which covers the overwhelming majority of protein coding genes in the human genome and a pool of 10 normal tissue samples as reference. Probesets were mapped to EnsEMBL genes.

Lee et al. [28] profiled Eyal-Giladi and Kochav Stage X embryos (a laid egg) on Affymetrix chicken genome gene arrays containing probes for all chicken coding genes. We eliminated probes not matching EnsEMBL V45 genes as well as probes mapping to more than one gene and averaged probes mapping to the same gene.

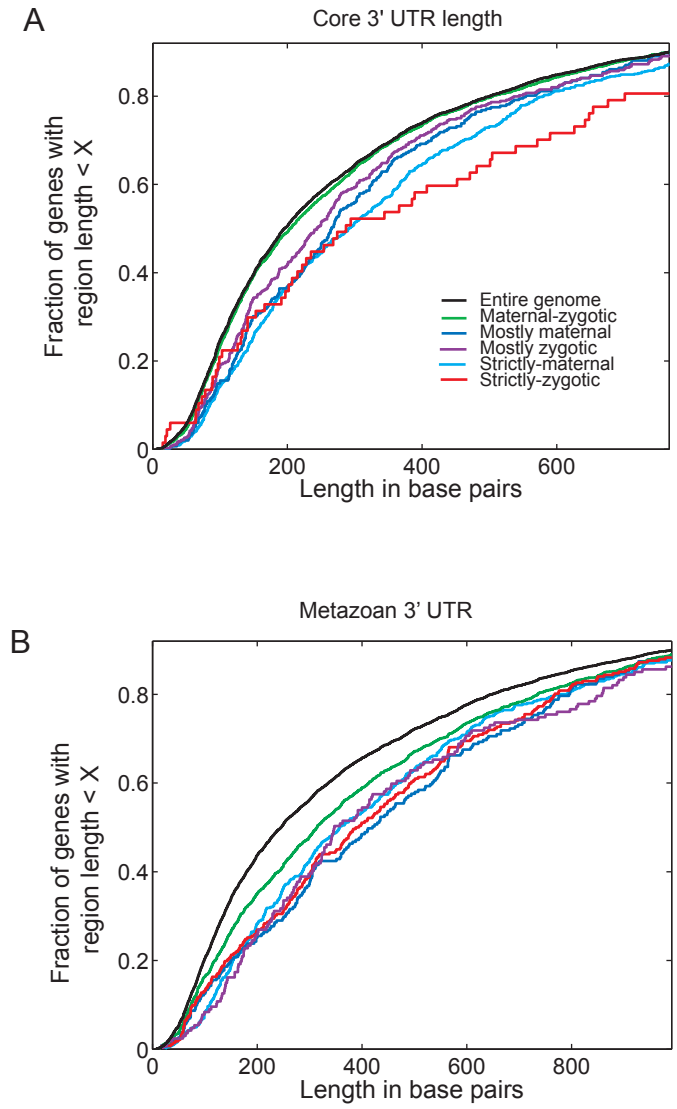

A
Core 3' UTR length
Fraction of genes with region length < X
0
0.2
0.4
0.6
0.8
0
200
400
600
Length in base pairs
Entire genome
Maternal-zygotic
Mostly maternal
Mostly zygotic
Strictly-maternal
Strictly-zygotic
B
Metazoan 3' UTR
Fraction of genes with region length < X
0
0.2
0.4
0.6
0.8
0
200
400
600
800
Length in base pairs

Figure S1.

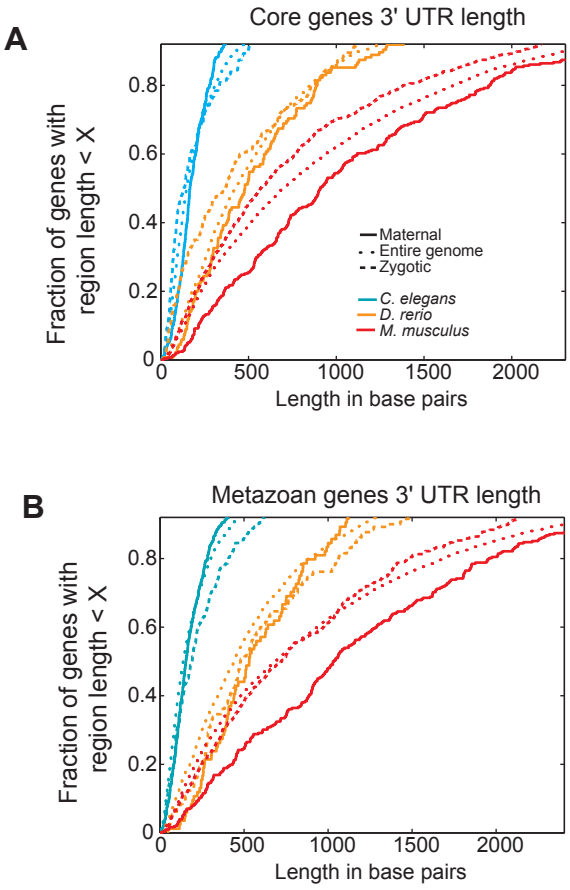

A
Core genes 3' UTR length
Fraction of genes with region length < X
0
0.2
0.4
0.6
0.8
0
500
1000
1500
2000
Length in base pairs
Maternal
Entire genome
Zygotic
C. elegans
D. rerio
M. musculus
B
Metazoan genes 3' UTR length
Fraction of genes with region length < X
0
0.2
0.4
0.6
0.8
0
500
1000
1500
2000
Length in base pairs

Figure S2.

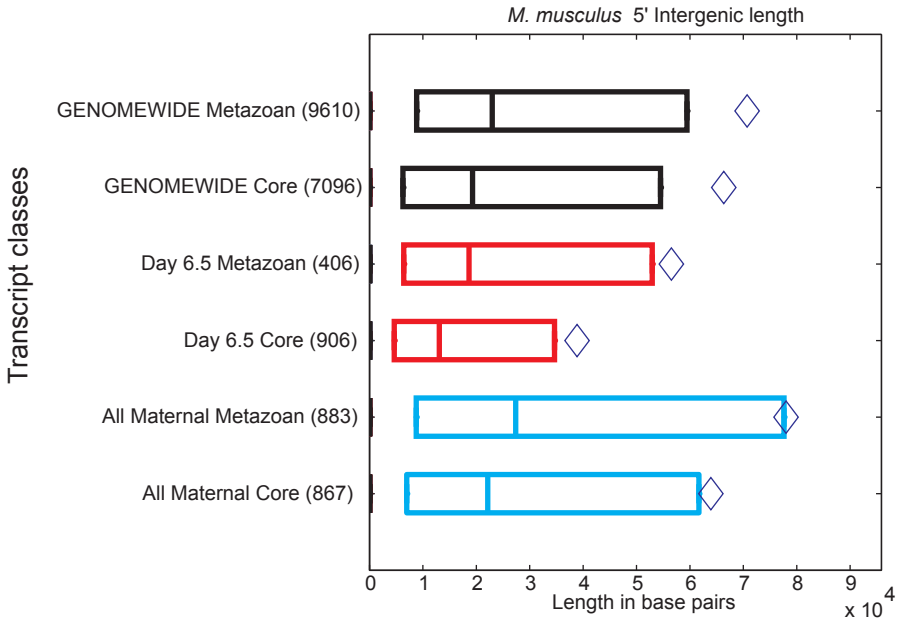
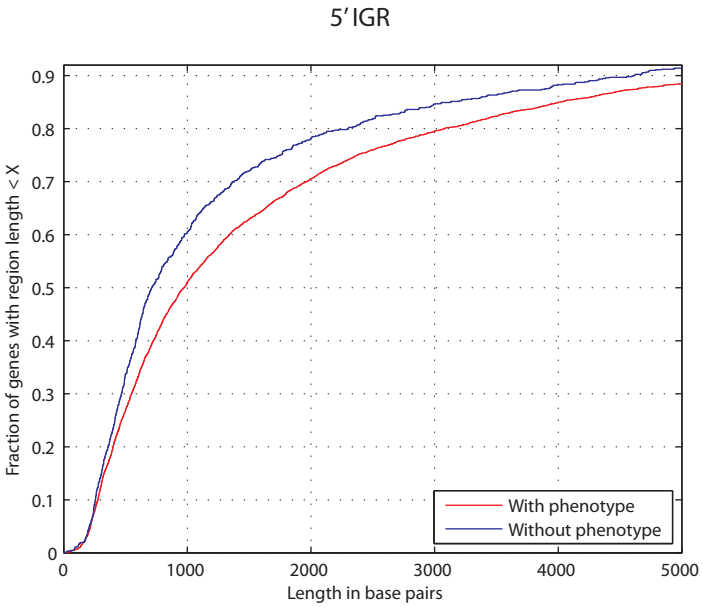

M. musculus 5' Intergenic length
Transcript classes
GENOMEWIDE Metazoan (9610)
GENOMEWIDE Core (7096)
Day 6.5 Metazoan (406)
Day 6.5 Core (906)
All Maternal Metazoan (883)
All Maternal Core (867)
0
1
2
3
4
5
6
7
8
9
Length in base pairs
x 10^4

Figure S3.

A


5′ IGR
Fraction of genes with region length < X
Length in base pairs
With phenotype
Without phenotype
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
0
1000
2000
3000
4000
5000

B

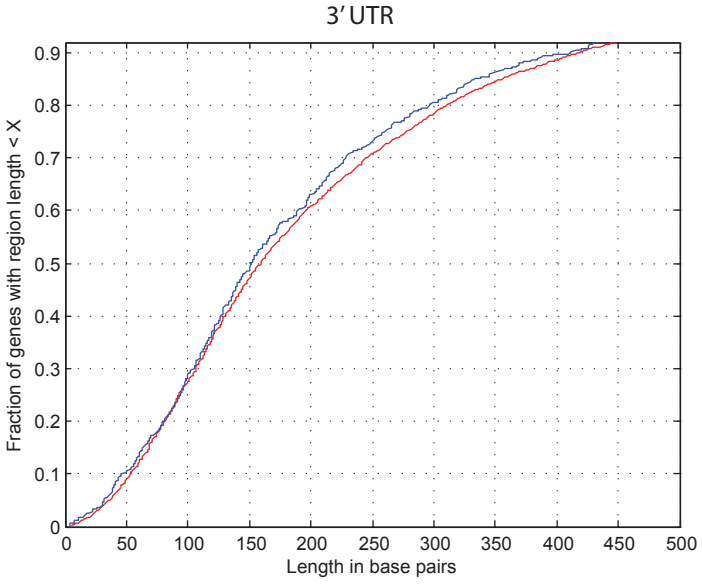
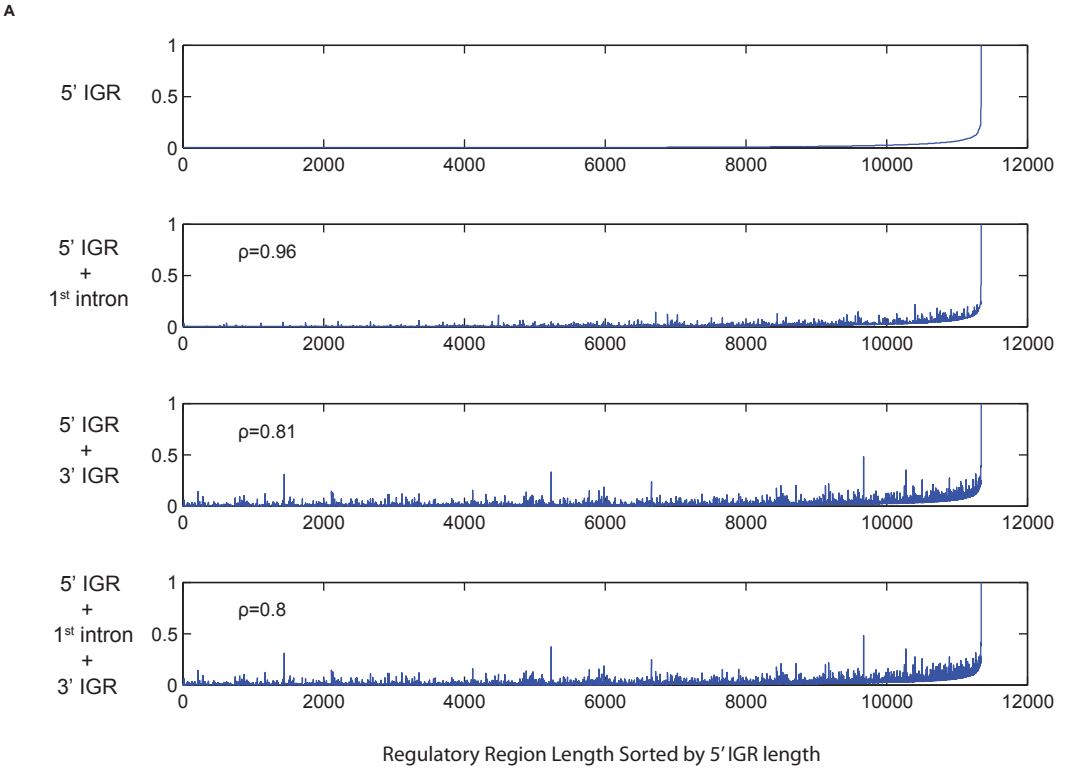

3′ UTR
Fraction of genes with region length < X
Length in base pairs
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
0
50
100
150
200
250
300
350
400
450
500

Figure S4.


A
5' IGR
5' IGR + 1st intron
ρ=0.96
5' IGR + 3' IGR
ρ=0.81
5' IGR + 1st intron + 3' IGR
ρ=0.8
0
0.5
1
0
2000
4000
6000
8000
10000
12000
Regulatory Region Length Sorted by 5' IGR length

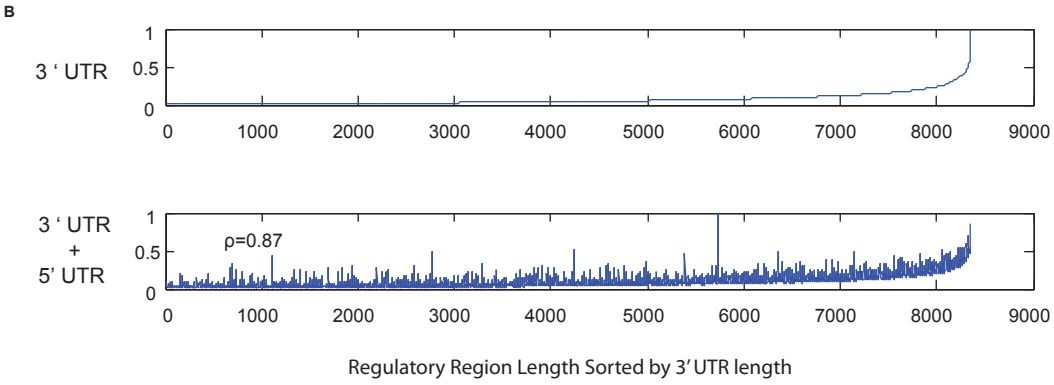

B
3 ' UTR
3 ' UTR + 5' UTR
ρ=0.87
0
0.5
1
0
1000
2000
3000
4000
5000
6000
7000
8000
9000
Regulatory Region Length Sorted by 3' UTR length

Figure S5.